DOCUMENT RESUME

ED 301 585 TM 012 456

AUTHOR Dungan, Lucille A.
TITLE Norm-Referenced Test Customization: Validation of Individual Score Interpretation.
PUB DATE Apr 88
NOTE 30p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (New Orleans, LA, April 6-8, 1988).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Achievement Tests; Difficulty Level; *Elementary School Students; Grade 4; Grade 6; Group Testing; *Individual Testing; Intermediate Grades; *Mathematics Tests; *Norm Referenced Tests; Sample Size; School Districts; Scoring; Test Construction; Test Interpretation; Test Items; *Test Validity; Urban Education
IDENTIFIERS *Metropolitan Achievement Tests; New York (New York Metropolitan Area); *Test Customization

ABSTRACT

This study investigated the "customization" of a standardized achievement test to measure the effect on group and individual scores. "Customization" refers to the modification of an existing test to more closely match test specifications of the entity using the test, usually a state or large school district. Results of an in-depth study of students in grades 4 and 6 in New York City are reported. Subject sample sizes ranged from 5,200 to 6,700 at grade 4 and from 6,200 to 7,600 at grade 6. Five hypothetically customized tests were constructed at each grade level by selecting five different General Survey Forms (sets of 20 piloted items) and substituting each set into a shelf test--the mathematics test of the Metropolitan Achievement Test, Sixth Edition (MAT6). This was done with the intention of creating the maximum difference between the difficulty of the shelf test and the customized New York test. Analysis of the effect on group and individual scores for the shelf test versus the customized test indicated that the differences occurred in a predictable direction. With the easiest items from the shelf test being removed and relatively difficult items substituted, the derived test was able to discriminate more at the high end of the MAT6 scale. Therefore, greater differences in scores for high scoring individuals were observed. The greatest changes in mean scaled scores were found in the problem solving subtest at both grade levels. (TJH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED301585

NORM-REFERENCED TEST CUSTOMIZATION:

VALIDATION OF INDIVIDUAL SCORE INTERPRETATION

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

JANE CANNER

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Lucille A. Dungan

The Psychological Corporation

A Paper Presented at the
Annual Meeting of the National Council on Measurement in Education
in New Orleans, Louisiana, April 8, 1988

TM 012 456

BEST COPY AVAILABLE

NORM-REFERENCED TEST CUSTOMIZATION:

VALIDATION OF INDIVIDUAL SCORE INTERPRETATION

The purpose of this study was to investigate the "customization" of a standardized achievement test to measure the effect on group and individual scores. Customization of a test refers to the modification of an existing test to more closely match test specifications of the entity using the new test (usually a state or large school district). Various methods used by states and school districts for customizing a test are discussed elsewhere (Keene and Holmes, 1987). Ultimately, the goal is to maintain a link to the national normative data of a major achievement test. The question, then, is how stable are group and individual scores between the original national shelf test and the newly customized test?

With the customization of a national achievement test, two types of changes are relevant to the issue of the validity of group and individual score interpretations. A customized test can be modified in terms of the content coverage of the items and in terms of the statistical properties of the items. With a customization process, the content coverage is intentionally modified to accommodate the testing of objectives not met by the original shelf test. The question of content change is a significant one because of the differential effect an item may have across individuals.

The statistical properties of test items and, thus, the statistical properties of the entire test, may be changed with the deletion of shelf test items and the insertion of new items. The most obvious characteristic subject to change is the difficulty of the test. If all new items are pretested (or piloted) with the national shelf test, item response theory can be used to calibrate all items on to one scale. This allows the customized test to be either pre-equated to the shelf test or changed in a desired direction (made more difficult or easier) based on the mean item difficulty of content areas and the total test.

The New York City Mathematics Testing Program provided data that allowed the investigation of changes in individuals' scores with the construction of hypothetically customized tests. In New York City, the complete Mathematics Test of the Metropolitan Achievement Test, Sixth Edition (MAT6) was administered to students in grades two through eight. Along with the MAT6 test, piloted items were administered in sets of twenty items (identified as General Survey forms) to samples of students at all grades. The piloted items consisted of newly written items and MAT6 Mathematics Diagnostic Test items selected by New York City personnel to match objectives not covered on the MAT6 test or to test material at a more advanced level than on the existing grade-level MAT6. With the pilot item data, it was possible to experiment with hypothetically constructing sample customized tests and measuring the effect on group and individual scaled scores.

This paper presents the results of an indepth study of two New York City grades, grades four and six. Five hypothetically customized tests were constructed at each grade level. The hypothetical New York City tests were built by selecting five different General Survey forms (sets of 20 piloted items) and substituting each set into the MAT6 shelf test. The 20 MAT6 items deleted were the easiest items within content areas. This was done with the intention of creating the maximum difference between the difficulty of the shelf test and the customized New York test. The effect on group and individual score differences between the shelf test and the newly customized tests was analyzed.

Methodology

Design

This study used the data from samples of New York City students in grades four and six who were tested on the Metropolitan Achievement Test, Sixth Edition (MAT6) Form M Mathematics Test in the spring of 1987. Sample sizes ranged from approximately 5200 to 6700 at grade four and 6200 to 7600 at grade six. Besides taking the MAT6 Mathematics Test, these groups of students took one of the New York City General Survey Forms (sets of pilot items). The General Survey Forms were distributed randomly throughout the City to obtain samples of students that would be representative of abilities of NYC students in general. This was done to enable the equating of all pilot items. With the use of item response theory for equating purposes, the samples did not need to be proportionally matched to all the characteristics of the entire City but simply representative of the full range of ability levels.

The General Survey Forms consisted of twenty items that were, on the average, more difficult than the average MAT6 item. This was particularly true for the two grades investigated in this study, four and six, where students were testing at the upper grade of a multiple grade-level MAT6 test.

Method

For purposes of analysis, hypothetical test forms were constructed at grades four and six where the greatest differences in item difficulty between MAT6 and General Survey items would be seen. General Survey sets of questions were substituted for twenty shelf test items. Substitutions were done within the three general content areas of the MAT6 Mathematics Test. These three areas are Concepts, Problem Solving, and Computation. The MAT6 Mathematics Test consists of a total of 95 items for the two levels investigated with the following breakdown by content areas:

MAT6 Mathematics Test Composition	
Concepts	35 items
Problem Solving	30 items
Computation	30 items

Each NYC General Survey form was analyzed for the number of items within the three content areas. Then, five General Survey forms were selected at each grade as replacements for MAT6 items. The forms selected contained the greatest percentage of items to be substituted into one content area, and, therefore, had the greatest potential to affect the difficulty level of one area of the newly customized test. The numbers of items substituted into each sample customized test are listed below.

Total Number of MAT6 Shelf Items and Number of Replacements

		Grade 4					Grade 6				
		General Survey Form					General Survey Form				
	MAT6	5	1	3	4	6	3	2	4	6	7
Concepts	35	9	5	10	7	9	8	12	10	10	8
Problem Solv..	30	7	11	6	10	6	10	8	8	8	10
Computation	30	4	4	4	3	5	2	0	2	2	2
Total	95	20	20	20	20	20	20	20	20	20	20

The General Survey items were substituted into the appropriate content areas of the shelf test with the corresponding number of MAT6 items being deleted. The MAT6 items replaced were the easiest items from each content area determined by item difficulty. For both choices, which General Survey forms were selected and which MAT6 items were deleted, the intention was to analyze the most extreme cases of item substitution and to analyze the degree of the resulting changes in group and individual scores.

All items from the General Survey forms were calibrated on to one scale and linked to the MAT6 scaled score using item response theory (Rasch model). The General Survey forms were equated using the common MAT6 items taken by all samples of students, and then all items were brought to the MAT6 scale using the appropriate forms and levels constants. Linking to the MAT6 scale allows the national test's norm base to be used.

Type of Analyses: Five General Survey forms were used at grades four and six to construct five hypothetically customized tests. The consistency of changes in group and individual scores across different sets of item replacements could then be investigated. For each customized test, three resulting test configurations and their associated scaled score distributions were analyzed. The possible test configurations are listed below.

		Number of Items
TEST CONFIGURATION 1	MAT6 (shelf)	K
TEST CONFIGURATION 2	MAT6-short (20 items deleted)	K-20
TEST CONFIGURATION 3	NY (20 items deleted plus 20 custom items)	K

Scores were obtained for each of the content areas (Concepts, Problem Solving, and Computation) and the total test for the three test configurations given above. The scaled scores used in this study are linear transformations of the log abilities produced for each raw score of a given test.

Group data were analyzed for changes in the mean scaled scores between the three test configurations, the MAT6, the MAT6-short, and the NY test. The magnitude of any differences between mean scaled scores were compared to the shift in mean item difficulty. Then, the variation in individual scaled scores between the three tests (MAT6, MAT6-short, and NY test) were looked at in several ways. Frequency distributions of the signed difference between the MAT6 and the NY test were generated to see the range of differences in individual scaled scores. The mean absolute differences between the three test configurations were also obtained. The comparisons made are displayed in the table below.

DIFFERENCE 1	MAT6 (shelf) - NY
DIFFERENCE 2	MAT6 (shelf) - MAT6 (short)
DIFFERENCE 3	NY - MAT6 (short)

Correlations between students' MAT6 scores and difference scores (MAT6-NY) were also obtained to see if there was a consistency in either underpredicting or overpredicting the MAT6 scores.

Results

Equating Samples

Based on the average mean MAT6 scaled scores for the samples of fourth and sixth grade students, there appears to be a fairly close match between sample means obtained in this study and citywide means. Both grades had mean MAT6 (shelf) scaled scores that were above average using individual norms. Grade six samples were larger in size and slightly more homogeneous than fourth grade samples.

Differences in Group Scores

Comparisons were made between the group mean scaled scores ($\bar{X}$ SS) for the three content areas (Concepts, Problem Solving, and Computation) and the total test for the three test configurations (MAT6, MAT6-short, and NY). These mean scaled scores for each of the General Survey Forms used as substitute items are listed in Table 1 for grade four and Table 2 for grade six. In addition to the mean scaled scores, each table gives the sample size (N), the number of items including the number of substituted items, and the mean item difficulties for the three test configurations. The mean item difficulties ($\bar{X}$ IDIFF) are the average of the Rasch difficulty estimates. Negative values indicate easier than average items, while positive values represent the more difficult items. To summarize the changes in mean scaled scores and mean item difficulties between the MAT6 (shelf) and NY tests, the differences (MAT6 - NY) are also listed in Tables 1 and 2. Most of the differences in mean scaled scores are negative indicating that the original MAT6 mean scaled score was greater than the newly derived NY mean scaled score. All of the mean item difficulties for the NY tests increase relative to the MAT6 mean item difficulties indicating the increased difficulty of the NY subtests and complete tests.

Overall, the greatest changes in mean scaled scores are seen in the Problem Solving subtest at both grades. Generally, the greatest percentage of substitutions and the greatest resulting changes in mean item difficulties took place within this content area. Looking at the data for all the subtests, as the difficulty of the test is increased from the shelf MAT6 to the derived NY test, a corresponding decrease is observed in the mean scaled score. There is a strong relationship between the degree of change in mean scaled scores and the magnitude of change in mean item difficulties.

Though there is a strong relationship between the change in mean item difficulty and the degree of the change in mean scaled score, this does not explain all the changes observed. In particular, though there is an increase in the difficulty of the Computation subtest going from the shelf to the NY tests, the mean scaled scores for the derived tests increase rather than decrease in four out of the five samples used. A suggested explanation for this difference will be discussed later.

Because of the varying direction of change across the three content areas, the total test statistics need to be analyzed separately. The degree of change observed in the total test mean scaled scores from shelf to NY tests is less than within subtests indicating that the variations in scores tend to cancel out across the entire test.

Looking at the group changes more closely, it can be seen that in some instances the change in mean scaled scores for a particular sample's subtest is greater than in other samples with an identical change in mean item difficulty. Clearly the change in mean item difficulty between the two tests is not the only statistical change affecting scores. The p values of the NY items (percentage of students answering a question correctly) substituted into the MAT6 were looked at in an attempt to explain anomalies between the magnitude of differences in scaled scores and item difficulties. While it was recognized that p values from various equating samples are not entirely comparable, the groups were homogeneous enough to allow the traditional item difficulty statistic to be used to indicate the relative difficulty of items.

The Computation subtests tended to have p values that were, overall, not only higher than the average p values of other subtests, but also had a smaller range of values. The Computation subtests that showed a resulting increase in mean scaled score going from shelf to NY test had no items with very low p values (i.e. less than .55). So even though the NY subtest was still relatively more difficult than the MAT6 subtest, the items being substituted into the new test were above average in difficulty and tended to have very few, if any, low p value items.

In analyzing the p values of the various sets of items substituted into the shelf test, it became clear that the ranges of p values were substantially different across different General Survey Forms. Though the mean p value of two sets of items could both be .50, one set of p values could range from .10 to .90 while a second set of p values could range from only .40 to .60. Compounding the differences in mean p values and ranges of p values, were differences in point biserial correlations. These discrimination indices also varied across items. These differences in the ranges of the p values and point biserials contribute to the differences in individual and group scores that cannot be seen from the magnitude of the change in mean item difficulties alone.

Differences in Individual Scores

In order to analyze differences in individual scaled scores between the original MAT6 shelf test scores and the derived test scores, both signed differences and absolute differences in scores between various test configurations were obtained. Tables 3 and 4 give the correlations between difference scores and the MAT6 scaled scores for grades four and six, respectively. First, the signed difference between individual MAT6 and NY scores (MAT6-NY) was correlated with the original MAT6 score to indicate the relationship between the direction of the change and the shelf test score. Secondly, the absolute difference (MAT6-NY) was correlated with MAT6 scores to indicate the relationship between the magnitude of the change across the MAT6 scale.

For the correlation coefficients with the individual signed difference scores, all the coefficients are positive. Since these difference scores range from high negative to high positive values, the positive correlations indicate that NY scaled scores tend to be higher than MAT6 scores at the low end of the MAT6 scale and NY scores are less than MAT6 scores for the high end of the MAT6 distribution of scaled scores. Derived NY test scores overpredict the MAT6 scores for low ability students while underpredicting scores for high ability students.

The magnitude of these correlation coefficients is significantly related to the degree of change in mean item difficulty between the MAT6 and NY tests. That is, the more the test difficulty is increased, the more the difference in individual scaled scores is related to position on the MAT6 scale. This indicates that with the method of item substitution used in this study (i.e. deleting the easiest items), as the test becomes increasingly difficult, greater variation is observed in scores at the tails of the distribution of MAT6 scores.

The correlation coefficients between the signed difference scores and MAT6 scores are usually higher for Concepts and Problem Solving than for Computation where the degree of difference in item difficulty and the percentage of item substitutions were greatest.

Also in Tables 3 and 4, correlations are given for the relationship between the absolute differences in scaled scores (MAT6 - NY) and MAT6 scores. Here we see a different effect than with the signed difference. Where there is a correlation (some coefficients are very close to zero), the coefficients are positive as was seen with the signed difference scores. This indicates greater changes in scores between shelf and NY tests as individuals score higher on the MAT6 scale. These coefficients tend to be greater for Computation than for Problem Solving and Concepts. The reverse was seen with the correlations for the signed differences. This seems to indicate that with less changes in the derived test (less increase in difficulty and fewer item substitutions), the greatest changes are occurring with higher scoring students. In some cases, a moderate correlation exists between MAT6 scores and both the signed difference and the absolute difference. This is probably due to the high degree of change in individuals with high scores. Plots of the difference scores versus MAT6 scores do indicate the greatest overall variation in individual scores is at the high end of the MAT6 scale.

To summarize overall differences in individual scaled scores, the next set of tables, Tables 5 and 6, gives means of the difference scores at grades four and six. The means of the signed differences between MAT6 and NY (MAT6-NY) scores for individuals are given first. These differences are fairly close to zero as would be expected from the similarity in group mean scaled scores between MAT6 and NY tests. As the group mean differences increase, mean differences in individual scaled scores increase so that we see the largest differences in the Problem Solving area and the smallest changes in Computation. The mean difference in individual scores for the total test tend to be smaller and more stable across derived test forms than the subtest differences as was seen with group mean scores. Most of the means of the signed differences in individual scores are positive indicating that MAT6 scores, on the average, are higher than the derived scores on the NY tests. In general, the customized NY tests that are more difficult than the original MAT6 shelf test produce estimates of ability (in terms of scaled scores) that are lower than would be obtained by actual MAT6 scores.

The mean absolute differences given next in Tables 5 and 6 indicate the overall unsigned variation in individual scores. Three comparisons are made in the tables and are summarized below.

ABDIFF1	MAT6 (shelf)	-	NY
ABDIFF2	MAT6 (shelf)	-	MAT6-short
ABDIFF3	NY	-	MAT6-short

A quick overview of the mean absolute differences in scaled scores indicate the greatest differences between MAT6 and NY test scores (ABDIFF1) and the smallest differences between the shelf and shortened MAT6 (ABDIFF2). The smallest differences would be expected to occur between the shelf and shortened MAT6 for several reasons. With the removal of the easiest items within a subtest, we are truncating the lower end of the ability estimate to raw score scale. Using one subtest as an example of this, the data for Concepts at grade four with General Survey Form 5 produces ability estimates ranging from -3.934 to 0.000 for raw scores on the MAT6 test of 1 through 15 (out of a total of 35 items). For the shortened MAT6 test, the ability estimates range from -2.760 to 0.000 for raw scores of 1 through 9 (out of 26 total items). So while the lowest ability estimates of the full shelf test are eliminated, this is the area of least discrimination for the majority of students, hence variations in scores are less than with changes in more difficult items.

Additionally, with the change from the shelf MAT6 to the shortened MAT6, all items are on the same original test (shelf) eliminating the need for an equating of items. Since no link is needed via an equating process (as was used with the General Survey Form items), there is one less source of variability in individual scores than in the MAT6 to NY comparison.

The mean absolute differences between the NY test and the shortened MAT6 test (ABDIFF3) are slightly smaller than the MAT6-NY comparison indicating that the direction of the change between the MAT6 and MAT6-short is the same as the direction of change between MAT6 and NY scores. All the mean absolute changes are very consistent across both grades four and six with the total test change again being less than subtests. For the total test score changes between MAT6 and NY (ABDIFF1), differences of between five and six scaled score points are observed at both grades.

Tables 7 and 8 are presented to indicate the ranges of signed differences in individual scaled scores between MAT6 and the NY tests (MAT6 - NY). Ranges are given for the middle sixty percent of the distribution (20-80%), the middle ninety percent (5-95%), and finally, the entire range of differences (0-100%). As was seen with the mean absolute differences in scores, the observed changes in individual scores are fairly consistent across both grades. The score differences at the 20-80% range indicate the range within which the majority of students scores varied. These values for the total test again are consistently around five scaled score points.

With small changes in the mean scaled scores differences (MAT6 - NY), the negative and positive ends of the ranges are approximately equal. With the increasing difficulty of a NY subtest (as in grade six Problem Solving subtests), the limit of the positive value of the range of differences increases relative to the negative limit. The reverse is seen with grade four Computation scores where mean scaled scores for the NY test increased relative to the MAT6 scaled score. Here we see higher negative limits of the ranges. This indicates that as difficult items are added to a test, there is a greater tendency for underpredicting of MAT6 scores and, if relatively easy items are added, more scores are overpredicted by the NY test.

While the potential for change in a single individual's score between the shelf and NY test is seen from the maximum range of differences (0-100%), those extreme cases are very rare. The maximum observed differences are significantly greater than the changes for the majority of individuals.

Discussion of Results

The purpose of this study was to maximize differences in individual scaled scores between a shelf test and a hypothetically customized test. The method used was to substitute custom items from New York City pilot items into the MAT6 shelf test. Sets of twenty custom items were substituted into appropriate content areas of the MAT6 shelf test by replacing the corresponding number of MAT6 items. The items deleted were the easiest MAT6 items within the given subtest. Scores were calculated for all individuals for three test configurations - MAT6-shelf, shortened MAT6 and derived NY test with the substituted custom items. Comparison of differences in group and individual scaled scores between the three test configurations were made to measure the effect of test customization with a significant change in the characteristics of the derived test.

The data from this study indicates that group means are stable even with the magnitude of change effected in this research. For the total NY test, across both fourth and sixth grades with ten different sets of substitutions, the greatest change in the mean scaled score was 1.5 scaled score points. This small change in the mean scaled score for the total test results, in part, from the cancellation of opposite effects from the three content areas.

The mean absolute differences between individual scores on the total mathematics tests (MAT6 - NY) all are between five and six scaled score points. Again, as was seen with the group mean scaled scores, the difference scores for the total test are of less magnitude and more consistent across forms than the differences for the three content areas with mean absolute differences ranging from six to 14 scaled score points.

The standard error of measurement for the total MAT6 Mathematics Test for students in the Spring national standardization sample is 3.8 raw score points at both grades four and six. These raw score points were translated into the scaled score metric at the approximate average scaled score of the New York samples used in this study. The standard error of 3.8 converts to approximately 12 scaled score units at both grades four and six.

Data in this study indicate mean differences of 5-6 scaled score points for total math with approximately 20% of the shelf test items being replaced. Since 80% of the test is held constant, differences due only to the reliability of the test would be substantially smaller than the 12 SEM scaled score points. Obtained differences of 5-6 points between actual MAT6 scores and the customized test scores are somewhat greater than would be expected due to test reliability alone.

The item statistics of substituted items, both difficulty and discrimination indices, are obviously important to changes in individuals' scores. But, of at least equal importance to the mean change (or stability) in test difficulty between a shelf and customized test, is the range of item difficulties being substituted into the new test form. The difficulty of an item determines where it will be most discriminating and, hence, where the greatest impact in score differences will be measured. Likewise, the individual characteristics of items substituted into a test appear to have a greater effect than simply the percentage of items replaced.

For the research in this paper, the entire set of twenty General Survey items replaced MAT6 items for the various hypothetical NY tests. Some of the pilot items had very low point biserial coefficients. (These were items that would not be selected for final forms of a test.) The low discriminating items could have an effect upon individual variation in scores. Ideally, to experiment with changes in scaled scores, it would be desirable to control these various statistical properties (mean p values, ranges of p values, and discrimination indices). However, the data limits us to the use of only the twenty items piloted within one equating sample. It becomes very difficult to control the item statistics in the desired manner with the limited sets of items available to us.

In general, the differences between shelf and customized test scores for samples of students in this study occurred in a predictable direction. With the easiest items from the shelf test being removed and relatively difficult items being substituted in, the derived test was able to discriminate more at the high end of the MAT6 scale and, therefore, greater differences in scores for high scoring individuals were observed. The direction of changes in scores reverses as one moves up the scaled score metric. At the low end of the MAT6 scale, the MAT6 is overpredicted by the derived test with the easy items removed. At the upper tail, the derived test score underpredicts the MAT6 scaled score.

Conclusions

Data presented in this paper provide evidence of the potential variation in group and individual scores when a shelf test is customized. The method for altering the standardized achievement test in this study involved increasing the difficulty of the test with the replacement of 21% of the total MAT6 items (20 out of 95). Substitutions within content areas were as high as 33%. The resulting changes in group and individual scores were analyzed.

Looking at the changes in group data between the shelf and customized test, differences in the mean scaled scores effected in this study for the large samples used were not substantial. Within content areas the maximum change in percentiles using individual norms was five points; the modal changes were only one or two percentile points. For the total test, the percentile corresponding to the mean scaled scores varied at most by one percentile point. These variations in percentile points would be expected to be greater using group norms. When using a group mean to make an interpretation from individual norms, the data are reliable with the amount of substitution taking place in this study.

Individual score changes between the shelf and customized tests were more variable than group data. Maximum changes in individual scores can be significant. However, the majority of students' scores were fluctuating within a general range of $\pm$ 6 to 10 scaled score points in the Concepts area, $\pm$ 9 to 16 points in Problem Solving, and $\pm$ 10 to 12 points in Computation. The total test score was less variable for individuals than the content area scores. For the total test score, average changes were $\pm$ 5 scaled score points. These changes must be considered in the context of the type of item replacement used in this study. The intention was to change the difficulty of the shelf test to the maximum degree possible with the custom items available for substitution. With a less severe change, individual scores would be expected to vary less.

In summary, as data are aggregated, the normative data become more precise. Group data are more stable than individual data; total test data are more stable than subtest scores.

Several differences between the method used in the current study and the actual New York City customization process must be noted. First, the pilot testing of New York items with the MAT6 test enabled the equating of the General Survey forms and the calibration of all items on to a common scale. Thus, with the development of New York City tests, it was possible to more closely match the mean item difficulty of the new forms to the shelf test's average difficulty than in this research project.

A second difference concerns the number of items replaced with the New York City customization. Final forms of the NYC Mathematics Test were expected to contain 50% of the original MAT6 shelf test items. This is a greater amount of substitution than was investigated in the current study.

In conclusion, the process of customizing a standardized achievement test can have differential effect across individuals. The types of changes expected to occur for individuals should be understood when norm-referenced score interpretations are to be made. With large group data, the individual variations in scores appear to cancel out across the distribution thus providing an accurate representation of average group performance relative to a national norm.

Reference

Keene, John M. and Holmes, Susan E., "Models for Obtaining Norm-Referenced Test Information for Statewide (Local) Objective-Referenced Tests: Issues and Challenges." A paper presented at the Annual Meeting of the National Council on Measurement in Education in Washington, D.C., April 22, 1987.

Table 1. Mean Scaled Scores and Mean Item Difficulties for MAT6 Shelf Tests, Shortened MAT6 Tests, and Hypothetical NY Customized Tests at Grade 4

General Survey Form	5			1			3		
N	6086			5958			5271		
	# Items/ # Subs[a]	$\bar{X}$ SS	$\bar{X}$ IDIFF	# Items # Subs[a]	$\bar{X}$ SS	$\bar{X}$ IDIFF	# Items/ # Subs[a]	$\bar{X}$ SS	$\bar{X}$ IDIFF
MAT6 CN[b]	35	620.6	.38	35	626.5	.38	35	623.1	.38
MAT6 CN-S[c]	26	620.3	.91	30	626.6	.71	25	622.6	.95
NY CN	35/9	619.8	.83	35/5	626.0	.72	35/10	621.2	.88
Difference[d]		-0.8	+.45		-0.5	+.34		-1.9	+.50
MAT6 PS	30	616.4	.18	30	622.2	.18	30	617.5	.18
MAT6 PS-S	23	616.0	.61	19	621.6	.86	24	617.3	.55
NY PS	30/7	615.4	.68	30/11	619.4	.89	30/6	616.2	.68
Difference		-1.0	+.50		-2.8	+.71		-1.3	+.50
MAT6 CP	30	604.9	-.63	30	606.4	-.63	30	605.3	-.63
MAT6 CP-S	26	604.9	-.32	26	606.4	-.32	26	605.4	-.32
NY CP	30/4	605.1	-.26	30/4	607.1	-.30	30/4	605.5	-.18
Difference		+0.2	+.37		+0.7	+.33		+0.2	+.45
MAT6 TOT	95	612.4	0.00	95	617.2	0.00	95	613.8	0.00
MAT6 TOT-S	75	612.4	.39	75	617.4	.39	75	613.6	.38
NY TOT	95/20	612.1	.44	95/20	616.5	.45	95/20	612.8	.48
Difference		-0.3	+.44		-0.7	+.45		-1.0	+.48

a Number of General Survey items substituted into MAT6 test

b CN - Concepts, PS - Problem Solving, CP - Computation, TOT - Total Test

c Shortened MAT6 test excluding easiest items

d MAT6 mean scaled score minus NY mean scaled score; NY mean item difficulty minus MAT6 mean item difficulty

Table 1. (continued)

General Survey Form	4			6		
N	6645			6704		
	# Items/ # Subs[a]	$\bar{X}$ SS	$\bar{X}$ IDIFF	# Items # Subs[a]	$\bar{X}$ SS	$\bar{X}$ IDIFF
MAT6 CN[b]	35	623.0	.38	35	615.1	.38
MAT6 CN-S[c]	28	622.9	.81	26	614.8	.91
NY CN	35/7	621.1	.88	35/9	614.6	.86
Difference[d]		-1.9	+.50		-0.5	+.48
MAT6 PS	30	617.5	.18	35	610.4	.18
MAT6 PS-S	20	616.3	.79	29	609.8	.55
NY PS	30/10	615.2	.95	35/6	607.4	.85
Difference		-2.3	+.77		-3.0	+.67
MAT6 CP	30	600.9	-.63	35	600.1	-.63
MAT6 CP-S	27	600.8	-.39	30	599.9	-.27
NY CP	30/3	601.4	-.20	35/5	598.7	-.15
Difference		+0.5	+.43		-1.4	+.48
MAT6 TOT	95	612.8	0.00	95	606.9	0.00
MAT6 TOT-S	75	612.5	.37	75	606.7	.40
NY TOT	95/20	611.8	.56	95/20	606.1	.54
Difference		-1.0	+.56		-0.8	+.54

Table 2. Mean Scaled Scores and Mean Item Difficulties for MAT6 Shelf Tests, Shortened MAT6 Tests, and Hypothetical NY Customized Tests at Grade 6

General Survey Form	3			2			4		
N	7140			6229			6853		
	# Items/ # Subs[a]	$\bar{X}$ SS	$\bar{X}$ IDIFF	# Items # Subs[a]	$\bar{X}$ SS	$\bar{X}$ IDIFF	# Items/ # Subs[a]	$\bar{X}$ SS	$\bar{X}$ IDIFF
MAT6 CN[b]	35	649.9	.52	35	650.7	.52	35	647.2	.52
MAT6 CN-S[c]	27	650.0	1.00	23	650.1	1.22	25	647.0	1.11
NY CN	35/8	649.9	.92	35/12	650.2	1.27	35/10	646.8	1.18
Difference[d]		+0.0	+.40		-0.5	+.75		-0.4	+.66
MAT6 PS	30	654.2	-.20	30	659.3	-.20	30	650.7	-.20
MAT6 PS-S	20	653.3	.38	22	658.6	.26	22	649.9	.26
NY PS	30/10	650.4	.74	30/8	653.8	.64	30/8	647.3	.61
Difference		-3.8	+.94		-5.5	+.84		-3.4	+.81
MAT6 CP	30	653.1	-.40	30	654.8	-.40	30	649.4	-.40
MAT6 CP-S	28	653.2	-.28	30	654.8	-.40	28	649.5	-.28
NY CP	30/2	652.7	-.14	30/0	654.8	-.40	30/2	647.8	-.13
Difference		-0.4	+.26		--	--		-1.6	+.27
MAT6 TOT	95	650.4	0.00	95	652.5	0.00	95	647.1	0.00
MAT6 TOT-S	75	650.3	.36	75	652.0	.29	75	646.8	.34
NY TOT	95/20	649.8	.53	95/20	651.0	.54	95/20	646.2	.59
Difference		-0.6	+.53		-1.5	+.54		-0.9	+.59

a Number of General Survey items substituted into MAT6 test

b CN - Concepts, PS - Problem Solving, CP - Computation, TOT - Total Test

c Shortened MAT6 test excluding easiest items

d MAT6 mean scaled score minus NY mean scaled score; NY mean item difficulty minus MAT6 mean item difficulty

Table 2. (continued)

General Survey Form	6			7		
N	7562			7459		
	# Items/ # Subs[a]	$\bar{X}$ SS	$\bar{X}$ IDIFF	# Items # Subs[a]	$\bar{X}$ SS	$\bar{X}$ IDIFF
MAT6 CN[b]	35	649.7	.52	35	650.2	.52
MAT6 CN-S[c]	25	649.6	1.11	27	650.1	1.00
NY CN	35/10	649.2	1.29	35/28	649.8	1.08
Difference[d]		-0.5	+.77		-0.4	+.56
MAT6 PS	30	653.7	-.20	30	655.9	-.20
MAT6 PS-S	22	652.9	.26	20	655.1	.38
NY PS	30/8	650.1	.46	30/10	650.6	.88
Difference		-3.6	+.66		-5.3	+1.08
MAT6 CP	30	648.7	-.40	30	651.6	-.40
MAT6 CP-S	28	648.7	-.28	28	651.6	-.28
NY CP	30/2	648.5	-.16	30/2	650.9	-.10
Difference		-0.2	+.24		-0.5	+.30
MAT6 TOT	95	648.9	0.00	95	650.3	0.00
MAT6 TOT-S	75	648.6	.34	75	650.1	.36
NY TOT	95/20	647.8	.57	95/20	649.2	.64
Difference		-1.1	+.57		-0.9	+.64

Table 3. Correlations between Individuals' MAT6 Scaled Scores and Difference Scores at Grade 4

MAT6 Scaled Scores correlated with the Signed Difference between MAT6 Scaled Score and NY Scaled Score

General Survey Form	5	1	3	4	6
CN	.39	.28	.55	.55	.32
PS	.17	.41	.31	.28	.57
CP	.02	.01	.04	.01	.25
TOT	.23	.34	.49	.41	.46

MAT6 Scaled Scores correlated with the Absolute Difference between MAT6 Scaled Score and NY Scaled Score

CN	.11	.05	.12	.28	-.07
PS	.27	.24	.28	.23	.40
CP	.29	.21	.44	.54	.44
TOT	.23	.23	.24	.28	.22

Table 4. Correlations between Individuals' MAT6 Scaled Scores and Difference Scores at Grade 6

MAT6 Scaled Scores correlated with the Signed Difference between MAT6 Scaled Score and NY Scaled Score

General Survey Form	3	2	4	6	7
CN	.18	.42	.35	.40	.38
PS	.49	.66	.45	.50	.62
CP	.10	—	.45	.14	.12
TOT	.39	.61	.47	.49	.54

MAT6 Scaled Score correlated with the Absolute Difference between MAT6 Scaled Score and NY Scaled Score

CN	-.05	.09	.02	.03	.10
PS	.29	.48	.34	.38	.37
CP	.54	—	.54	.52	.56
TOT	.22	.38	.25	.25	.27

Table 5. Group Means of Difference Scores at Grade 4

General Survey Form		5	1	3	4	6
DIFF1[a]	CN	.84	.51	1.83	1.89	0.42
	PS	.99	2.76	1.28	2.35	3.07
	CP	-.12	-0.71	-0.17	-0.53	1.38
	TOT	.33	0.71	1.06	1.01	0.77
ABDIFF1[b]	CN	8.59	6.39	9.54	8.08	8.32
	PS	9.29	13.75	8.98	12.07	9.94
	CP	7.37	6.78	8.36	8.21	9.18
	TOT	5.11	5.85	5.44	5.61	5.48
ABDIFF2[c]	CN	4.97	2.44	5.43	3.53	5.20
	PS	5.08	7.91	4.32	7.40	4.51
	CP	1.64	1.68	1.72	1.33	2.51
	TOT	2.85	2.90	2.86	2.96	2.84
ABDIFF3[d]	CN	7.56	5.75	8.28	7.17	7.12
	PS	8.40	12.70	8.09	10.90	8.99
	CP	7.02	6.40	7.96	7.86	8.74
	TOT	4.48	5.17	4.83	4.99	4.82

a Mean signed difference of MAT6 scaled score minus NY scaled score
b Mean absolute difference of MAT6 scaled score minus NY scaled score
c Mean absolute difference of MAT6 scaled score minus shortened MAT6 scaled score
d Mean absolute difference of NY scaled score minus shortened MAT6 scaled score

Table 6. Group Means of Difference Scores at Grade 6

General Survey Form		3	2	4	6	7
DIFF1[a]	CN	-0.09	0.52	0.38	0.42	0.40
	PS	3.72	5.50	3.49	3.70	5.27
	CP	0.48	--	1.61	0.28	0.67
	TOT	0.61	1.44	0.95	1.07	1.11
ABDIFF1[b]	CN	7.07	9.32	8.95	8.64	7.42
	PS	13.83	13.55	11.81	12.05	14.11
	CP	6.86	--	7.49	5.95	6.56
	TOT	5.69	6.06	6.03	5.52	5.69
ABDIFF2[c]	CN	4.43	6.26	5.37	5.24	4.18
	PS	7.12	5.38	5.62	5.53	6.95
	CP	1.02	--	1.11	1.04	1.00
	TOT	2.93	3.13	2.82	2.78	2.78
ABDIFF3[d]	CN	6.36	8.41	7.81	7.12	6.29
	PS	13.13	13.39	11.16	11.64	13.87
	CP	6.53	--	7.31	5.86	6.11
	TOT	4.98	5.62	5.49	4.84	5.05

a Mean signed difference of MAT6 scaled score minus NY scaled score
b Mean absolute difference of MAT6 scaled score minus NY scaled score
c Mean absolute difference of MAT6 scaled score minus shortened MAT6 scaled score
d Mean absolute difference of NY scaled score minus shortened MAT6 scaled score

Table 7. Ranges of Differences in Individual Scaled Scores (MAT6 - NY) for 60%, 90% and 100% of students at Grade 4

General Survey Form		5		1		3		4		6	
20-80%	CN	-9	10	-6	7	-8	12	-6	9	-8	9
	PS	-10	11	-12	15	-8	10	-9	15	-6	10
	CP	-8	7	-7	5	-12	8	-13	6	-8	10
	TOT	-5	5	-5	6	-4	6	-5	6	-5	6
5-95%	CN	-16	18	-11	13	-19	20	-14	18	-18	17
	PS	-16	18	-24	32	-16	19	-24	27	-17	27
	CP	-10	19	-9	16	-13	15	-14	15	-14	19
	TOT	-10	10	-11	13	-10	12	-11	13	-11	12
0-100%	CN	-76	73	-43	52	-58	66	-56	58	-63	48
	PS	-58	69	-70	98	-60	69	-81	80	-51	71
	CP	-58	50	-36	51	-39	58	-40	56	-41	69
	TOT	-51	66	-38	66	-39	58	-41	71	-40	67

Table 8. Ranges of Differences in Individual Scaled Scores (MAT6 - NY) for 60%, 90% and 100% of students at Grade 6

General Survey Form		3		2		4		6		7	
20-80%	CN	-7	7	-9	10	-9	8	-8	9	-7	7
	PS	-11	16	-8	17	-9	15	-9	14	-8	17
	CP	-10	6	-	-	-9	7	-8	6	-5	5
	TOT	-6	6	-5	7	-5	6	-5	6	-5	6
5-95%	CN	-15	15	-19	19	-18	19	-19	17	-15	14
	PS	-24	33	-9	40	-22	30	-19	32	-22	36
	CP	-10	15	-	-	-11	16	-10	15	-16	12
	TOT	-12	12	-10	15	-11	15	-10	12	-11	12
0-100%	CN	-77	65	-51	70	-52	81	-61	58	-42	66
	PS	-83	79	-55	78	-81	80	-64	79	-81	88
	CP	-37	42	-	-	-38	41	-37	43	-43	37
	TOT	-52	54	-29	88	-33	70	-34	79	-43	70